

Alexander Kopleinig, Peter Meyer, Carolin Müller-Spitzer

Dictionary users do look up frequent words. A log file analysis

Abstract: In this paper, we use the 2012 log files of two German online dictionaries (Digital Dictionary of the German Language¹ and the German version of Wiktionary) and the 100,000 most frequent words in the Mannheim German Reference Corpus from 2009 to answer the question of whether dictionary users really do look up frequent words, first asked by de Schryver et al. (2006). By using an approach to the comparison of log files and corpus data which is completely different from that of the aforementioned authors, we provide empirical evidence that indicates – contrary to the results of de Schryver et al. and Verlinde/Binon (2010) – that the corpus frequency of a word can indeed be an important factor in determining what online dictionary users look up. Finally, we incorporate word class information readily available in Wiktionary into our analysis to improve our results considerably.

Keywords: log file, frequency, corpus, headword list, monolingual dictionary, multi-lingual dictionary

Alexander Kopleinig: Institut für Deutsche Sprache, R 5, 6-13, 68161 Mannheim, +49-(0)621-1581-435, kopleinig@ids-mannheim.de

Peter Meyer: Institut für Deutsche Sprache, R 5, 6-13, 68161 Mannheim, +49-(0)621-1581-427, meyer@ids-mannheim.de

Carolin Müller-Spitzer: Institut für Deutsche Sprache, R 5, 6-13, 68161 Mannheim, +49-(0)621-1581-429, mueller-spitzer@ids-mannheim.de

Introduction

We would like to start this chapter by asking one of the most fundamental questions for any general lexicographical endeavour to describe the words of one (or more) language(s): which words should be included in a dictionary? At first glance, the answer seems rather simple (especially when the primary objective is to describe a language as completely as possible): it would be best to include every word in the dictionary. Things are not that simple, though. Looking at the character string ‚bfk‘, many people would probably agree that this ‘word’ should not be included in the dictionary, because they have never heard anyone using it. In fact, it is not even a

¹ We are very grateful to the DWDS team for providing us with their log files.

word. At the same time, if we look up “afk” in Wiktionary², a word that many people will not have ever heard or read, either, we find that it is an abbreviation that means *away from (the computer) keyboard*. In fact, as we will show below, “afk” was one of the 50 most looked-up words in the German version of Wiktionary in 2012. So, maybe a better way to answer the question of which words to include in the dictionary is to assume that it has something to do with usage. If we consult official comments about five different online dictionaries, this turns out to be a wide-spread assumption:

“How does a word get into a Merriam-Webster dictionary? This is one of the questions Merriam-Webster editors are most often asked. The answer is simple: usage.”³

“How do you decide whether a new word should be included in an Oxford dictionary? [...] We continually monitor the Corpus and the Reading Programme to track new words coming into the language: when we have evidence of a new term being used in a variety of different sources (not just by one writer) it becomes a candidate for inclusion in one of our dictionaries.”⁴

„Die Erzeugung der *lexiko*-Stichwortliste erfolgte im Wesentlichen in zwei Schritten: Zunächst wurden die im Korpus vorkommenden Wortformen auf entsprechende Grundformen zurückgeführt; diese wurden ab einer bestimmten Vorkommenshäufigkeit in die Liste der Stichwortkandidaten aufgenommen.“⁵ [‘The *lexiko* headword list was essentially created in two steps: first of all, the word forms which occurred in the corpus were reduced to their respective basic forms; and then those that attained a particular frequency of occurrence were included in the list of headword candidates.’]

„Wie kommt ein Wort in den Duden? Das wichtigste Verfahren der Dudenredaktion besteht darin, dass sie mithilfe von Computerprogrammen sehr große Mengen an elektronischen Texten daraufhin „durchkämmt“, ob in ihnen bislang unbekannte Wörter enthalten sind. Treten sie in einer gewissen Häufung und einer bestimmten Streuung über die Texte hinweg auf, handelt es sich um Neuaufnahmekandidaten für die Wörterbücher.“⁶ [‘How does a word get into the Duden? The most important process carried out by the Duden editors consists of using computer programs to „comb through“ large quantities of electronic texts to see whether they contain words which were previously unknown to them. If they appear across the texts in particular numbers and in a particular distribution, then they become new candidates for inclusion in the dictionaries.’]

“Some Criteria for Inclusion [...] Frequency: The editors look at large balanced, representative databases of English to establish how frequently a particular word occurs in the language.

² <http://en.wiktionary.org/wiki/AFK> (last accessed 20 June 2013).

³ http://www.merriam-webster.com/help/faq/words_in.htm?&t=1371645777 (last accessed 20 June 2013).

⁴ <http://oxforddictionaries.com/words/how-do-you-decide-whether-a-new-word-should-be-included-in-an-oxford-dictionary> (last accessed 20 June 2013).

⁵ <http://www1.ids-mannheim.de/lexik/elexiko/methoden.html> (last accessed 20 June 2013).

⁶ http://www.duden.de/ueber_duden/wie-kommt-ein-wort-in-den-duden (last accessed 20 June 2013).

Words that do not occur in these databases, or only occur with a minuscule frequency, are not likely to be included in the dictionary.”⁷

Thus, one essential requirement for a word to be included in the dictionary is usage. Of course, it is an enormous (or maybe impossible) project to include every word in the dictionary that is used in the language in question. Even in the case of electronic dictionaries which do not share the natural space limitations of their printed counterparts, the fact must be faced that writing dictionary entries is time-consuming and labour-intensive, so every dictionary compiler has to decide which words to include and just as importantly which words to leave out. The last four of the five statements quoted above show how lexicographers often solve this problem practically. The answer is, of course, frequency of use which is measured using a corpus. Only if the frequency of a word exceeds a (rather arbitrarily) defined threshold does it then become a candidate for inclusion in the dictionary. Again, for most lexicographical projects, this definition turns out to be problematic. What if more words exceed this frequency threshold than could be described appropriately in the dictionary given a limited amount of time and manpower? In this case, the threshold could just be raised accordingly. However, this again just means that it is implicitly assumed that it is somehow more important to include more frequent words instead of less frequent words.

In this chapter, we would like to tackle this research question by analyzing the log files of two German online dictionaries. Does it actually make sense to select words based on frequency considerations, or, in other words, is it a reasonable strategy to prefer words that are more frequent over words that are not so frequent? Answering this question is especially important when it comes to building up a completely new general dictionary from scratch and the lexicographer has to compile a headword list, because if the answer to this question was negative, lexicographers would have to find other criteria for the inclusion of words in their dictionary.

The rest of this chapter is structured as follows: in the next section, we review previous research on the analysis of log files with regard to the question just outlined; in Sections 3 and 4, we summarize how we obtained and prepared the data that are the basis of our study and that is described in Section 5; Section 6 focuses on our approach to analyzing the data, while Section 7 ends this chapter with some concluding remarks.

⁷ <http://www.collinsdictionary.com/words-and-language/blog/collins-dictionary-some-criteria-for-inclusion,55,HCB.html> (last accessed 20 June 2013).

1 Previous research

To understand whether including words based on frequency of usage considerations makes sense, it is a reasonable strategy to check whether dictionary users actually look up frequent words. Of course, in this specific case, it is not possible to design a survey (or an experiment) and ask potential users whether they prefer to look up frequent words or something like that. That is why de Schryver and his colleagues (2006) conducted an analysis where they compared a corpus frequency list with a frequency list obtained from log files. Essentially, log files record, among other things, search queries entered by users into the search bar of a dictionary. By aggregating all individual queries, it is easy to create a frequency list that can be sorted just like any other word frequency list. The aim of de Schryver et al.'s study was to find out if dictionary users look up frequent words, because:

“it seems as if treating just the top-frequent orthographic words in a dictionary will indeed satisfy most users, and this in turn seems to indicate that a corpus-based approach to the macro-structural treatment of the 'words' of a language is an excellent strategy. This conclusion, however, is *not* correct, as will be shown” (de Schryver et al., 2006, p. 73, emphasis in original)

To analyze their data, de Schryver et al. correlated the ranked corpus frequency with the ranked look up frequency. Statistically speaking, correlation refers to the (linear) relationship between two given variables, which is just a scale-independent version of the covariance of those two variables. Covariance measures how two variables x and y change together: if greater values, i.e. values above average, of x mainly correspond with greater values of y , it assumes positive values. By dividing the covariance by the product of the respective standard deviations, we obtain a scale-independent measure ranging from -1 to 1 (cf. Ludwig-Mayerhofer, 2011). It is important to emphasize that a strong correlation also implies that smaller values of x mainly correspond to smaller values of y . Therefore the question that de Schryver et al (2006) actually tried to answer is: do dictionary users look up *frequent* words *frequently*? And, do dictionary users look up *less frequent* words *less frequently*? The result of their study is part of the title of their paper: “On the Overestimation of the Value of Corpus-based Lexicography”. Verlinde & Binon (2010, p. 1148) replicated the study of de Schryver et al. (2006) using the same methodological approach and essentially came to the same conclusion.

In Section 4, we will try to show why de Schryver et al.'s straightforward approach is rather problematic due to the distribution of the linguistic data that are used. In this context we suggest a completely different approach and show that dictionary users do indeed look up frequent words (sometimes even frequently). This is why we believe that dictionary compilers do not overestimate the value of corpus-based lexicography.

2 Obtaining the data

All log file and corpus input data for our study are represented in plain text files with a simple line-based character-separated (CSV) format. Each line consists of a character string representing a word, sequence of words, or query string, followed by a fixed delimiter string and further information on the character string, typically a number representing the token frequency of that string in a corpus or the number of lookups in a specific dictionary. The following sections present a brief overview of how the various files were obtained or generated, including some technical details for interested readers.

Corpus data

As a corpus list, we used an unpublished version of the unlemmatised DEREW0 list which contains the 100,000 most frequent word forms in the Mannheim German Reference Corpus (DEREKO) paired with their respective raw frequencies. DEREKO is “one of the major resources worldwide for the study of the German language” (Kupietz, Belica, Keibel, & Witt, 2010, p. 1848).⁸

The dictionaries

Both the Digital Dictionary of the German Language (DWDS) and the German version of Wiktionary are general dictionaries that do not describe specialized vocabulary for a specific user group, but endeavour to describe the vocabulary of German as comprehensively as possible. The DWDS is a monolingual dictionary project which tries to bring together and update the available lexical knowledge that can be found in existing comprehensive dictionaries⁹. The German version of Wiktionary is a multilingual dictionary (Meyer & Gurevych, 2012) which also focuses on the description of the German vocabulary as a whole and is freely available for the general public.¹⁰

The DWDS and Wiktionary are suitable dictionaries for the research question presented above for the following reasons:

- Both dictionaries have a broad scope. Therefore, a diverse consultation behaviour regarding German vocabulary can be expected. That is why the log file data

⁸ We used the most recent version of this list published in May 2009 available here <http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html> (last accessed 25 June 2013). Instead of raw frequencies, this list only contains frequency classes (cf. the user documentation for further details); we thank our colleague Rainer Perkuhn for providing us with the respective raw frequencies.

⁹ <http://www.dwds.de/projekt/hintergrund/> (last accessed 25 June 2013).

¹⁰ http://de.wiktionary.org/wiki/Wiktionary:%C3%9Cber_das_Wiktionary (last accessed 25 June 2013).

can be used to check whether users really do look up words that are frequent in a corpus.

- Both dictionaries are used frequently, so it is rather unlikely that particular special search requests will bias the data.¹¹

The fact that Wiktionary is based on user-generated content is not a problem for our purposes, because most of the criticisms in the context of Wiktionary are not in any case directed at the coverage of terms (which is very broad, as we will show below), but at the structure of the entries which in many cases either is outdated, does not take into account current lexicographical research or presents insufficient source and usage information (Hanks, 2012, pp. 77–82; Nesi, 2012, pp. 373–374; Rundell, 2012, pp. 80–81).

DWDS log files

We processed the log files generated by the DWDS web application between January 28, 2012, and January 8, 2013. The files have a simple standard line-based plain text format, with each line representing one HTTP request and specifying, amongst other things, the IP address of the HTTP client, the exact time of the request, and the so-called HTTP request line that contains the URI of the requested resource. A Java program processed all log files using regular expressions, selecting all requests representing the action of looking up a word (or, more generally, a character string) in any of the presentation modes offered by the DWDS web portal. This includes all cases where the lookup process was initiated by following a hyperlink, i.e., the HTTP referer was not taken into account. In order to comply with standard privacy policies, IP addresses were bijectively mapped onto arbitrary integers. A simple character code was used to indicate private IP addresses. The resulting intermediate CSV file has a size of 160.5 MB and contains 3,366,426 entry lines of the following format:

```
-|1234|29/Apr/2012:06:48:54 +0200|Herk%C3%B6mmlich
```

This sample line indicates that a request to look up the string ‘Herkömmlich’ in the German Wiktionary was issued on April 29 from the IP address with serial number 1234. The lookup string is represented in URL-encoded format in the log files; the IP address is from a public address space as indicated by the initial ‘-’.

Secondly, a script written in Groovy¹² processed the intermediate CSV file by removing the URL encoding and counting all occurrences of each query string contained in the logs. The resulting CSV file contains 581,283 lines, i.e., the DWDS log

¹¹ This was also the reason why we did not use the log files of one of the IDS dictionaries, since all of those dictionaries are either specialized or not consulted frequently enough.

¹² See <http://groovy.codehaus.org> (last accessed 20 June 2013).

files of almost a complete year register more than half a million different query strings.

Wiktionary log files

The Wikimedia Foundation¹³ publishes hourly page view statistics log files where all requests of any page belonging to one of the projects of the Foundation (such as Wikipedia, Wiktionary and others) within a particular hour are registered. Each log file entry indicates the title of the page retrieved, the name of the Wikimedia project the page belongs to, the number of requests for that page within the hour in question, and the size of the page's content. Request figures are not unique visit counts, i.e., multiple requests of a page from the same IP address are treated as distinct page views.

We used a Groovy script to analyze all page view files from the year 2012. For each month, there is a separate index page¹⁴ containing links to all gzip-compressed hourly log files of that month. Our script follows all of the roughly 700 links of each index page. Reading in the contents of the URL, decompressing them and parsing them line by line is performed in memory using a chain of standard Java input streams. This keeps the memory and hard disk footprint for processing more than 2.5 terabytes of plain text data to a minimum, the only remaining bottleneck being network bandwidth.

The script scans each of the 8,784 hourly log files for entries concerning regular article pages in the German Wiktionary (which is the project resource indicated by a line-initial “de.d” in the log), irrespective of whether the requested page title is in German or any other language. There is a sum total of 91,271,569 such entries; the request counts for each page title found were added together and written to a CSV file that contains 1,621,249 entries.¹⁵

Wiktionary word class information

The Wikimedia foundation publishes complete dumps of all data of its projects at regular intervals. We used a bzip2-compressed XML dump file of the current text and metadata of the pages of the German Wiktionary on June 3, 2013,¹⁶ as the basis for a rough-and-ready mapping of words onto word class information in a wide

¹³ Cf. <http://wikimediafoundation.org> (last accessed 20 June 2013).

¹⁴ The index page URL is <http://dumps.wikimedia.org/other/pagecounts-raw/2012/2012-mm; mm = 01...12>. (last accessed 20 June 2013).

¹⁵ For practical reasons, any page that was viewed only once within a whole month was discarded from the statistics for that month. This procedure reduces the number of pages to consider to less than a quarter. The lookup frequency of such rare page views is far below the threshold we chose for our analysis.

¹⁶ The download URL for the file is <http://dumps.wikimedia.org/dewiktionary/20130603/dewiktionary-20130603-pages-articles.xml.bz2> (last accessed 20 June 2013).

sense, including a classification of word forms as first or last name, toponym, or inflected form. The uncompressed size of the dump file is about 450 MB. In the XML document, each Wiktionary page is represented by a <page> element that contains metadata and the content proper in a Wikimedia-specific markup format.¹⁷ We analyzed the XML file with a standard Java-based SAX parser, using a regular expression to extract all ‘part of speech’ header information from the different sections of the markup of each page. The results were written into a CSV file pairing the 123,578 page titles with the sequence of all ‘part of speech’ classifications for the page in question. The remaining 146,705 pages contained in the dump do not contain any ‘part of speech’ headers.

3 Preparing the data

Corpus data

To make the different sets of data intercomparable, we first replaced all word forms in the DEREWO list with their lowercase variant.¹⁸ After this, the frequencies of duplicate word forms were added together¹⁹ and each word form received a rank according to its raw frequency. One caveat is in order here: there are of course word forms that have the same frequency.²⁰ Thus, a decision has to be made as to how to rank these word forms. There are several possibilities, for example generating average ranks for all word forms with an identical raw frequency count. However, we opted for a rather pragmatic procedure: word forms with identical frequencies were ranked randomly, because (contrary to de Schryver et al.’s approach) this does not make any difference to the results of our analysis, as will be shown below. In total, we generated a list with the 92,506 most frequent DEREKO word forms.

DWDS & Wiktionary log files

As mentioned above, we were primarily interested in a comparison between the log files and the DEREWO list, and not in the question of what users generally look for. Since the corpus list only consists of unigrams, we first removed all n-grams with $n > 1$ from the log files. Furthermore, we removed queries that were longer than 120

¹⁷ See, e.g., http://en.wikipedia.org/wiki/Help:Wiki_markup (last accessed 20 June 2013).

¹⁸ This is an important step, because many users of electronic dictionaries assume that the search function is case insensitive, so they pay no attention to capitalization.

¹⁹ For example the German definite masculine article “der” appeared both in its lowercase version and in the uppercase one. “der” has a raw frequency of 109,354,718, while “Der” has a frequency of 12,926,941, so after the data preparation, “der” is listed in the data with an adjusted frequency of 122,281,659.

²⁰ Actually only 28.15% of the word forms have a unique frequency.

characters or queries containing numbers and special characters.²¹ While we admit that these steps are worthy of discussion, we believe that this procedure again is necessitated by the (unigram) structure of the DEREWO list. Furthermore, additional calculations show that those steps only remove 4.8% of the DWDS and 7.4% of the Wiktionary raw log file tokens.

The resulting lists were then prepared in the same way as the corpus data. In total, we generated a list with 1,287,365 Wiktionary log file types and a list with 156,478 DWDS log file types.

4 Describing the data

Corpus data

If we look at the DEREWO list and plot the relative frequency against the rank, we receive a typical Zipfian pattern (cf. Fig 1a for the first 1,000 ranks). This means that we have a handful of word forms that have a very high frequency and an overwhelming majority of word forms that have a very low frequency. Or, in other words, our DEREWO list consists of 3,227,479,836 word form tokens. The 200 most frequent word form types in the list make exactly half of those tokens.

Log files

As mentioned in the previous section, the Wiktionary log file types are roughly 8 times as big as the DWDS log file types. To make the results both comparable and more intuitive, we rescaled the data by multiplying the raw frequency of a query by 1,000,000, dividing it by the sum of all query tokens and rounding the resulting value. We then removed all queries with a value smaller than one.²² Thus, the result-

²¹ $\frac{3}{4}$, \varnothing , z , Υ , $\text{\textcircled{a}}$, i , TM , \llcorner , \textcent , \pm , \backslash , $\#$, $!$, $\$$, $/$, $\&$, \cdot , $\text{\textcircled{a}}$, $\text{\textcircled{c}}$, $($, $\%$, $)$, $*$, $+$, $;$, $<$, $>$, $?$, $=$, $[$, $]$, \wedge and search requests starting with a hyphen.

²² We think that the scaling is an important step to make the results of the different log file sources intercomparable. Please note that a value smaller than 1 means that the string in question is searched for less than 0.5 times in 1 million search requests. For the DWDS log files, no data were dropped. For the Wiktionary log files, this procedure dropped 4.4 % of all 2012 search request tokens, which, due to the distribution of the frequency list, amounts to 85.6% of all search types. So we only used the remaining 14.4 % for our analyses. Nevertheless, as can be seen in Table 1, we still used the first 185,071 most frequent Wiktionary search request types for the analyses, which is more than all 2012 DWDS search request types. Furthermore, the DWDS and the Wiktionary data point in the same direction, which makes it rather unlikely that the effects we describe are only artefacts resulting from this step. However, to make sure that this step was not a problem for our conclusions, we reran all analyses presented in this contribution for the Wiktionary data without removing any search requests (with values smaller than 1 replaced by 1), so that no data were dropped. In general, those analyses show that the scaling of the data does not invalidate any conclusions drawn; only the Wiktionary token figures presented in Table 8 are smaller for the regular searches, of course.

ing variable is measured in a unit that we would like to call *poms*. For example, a value of 8 means that the corresponding phrase is searched for 8 times *per one million* search requests. Table 1 summarizes the resulting distribution.

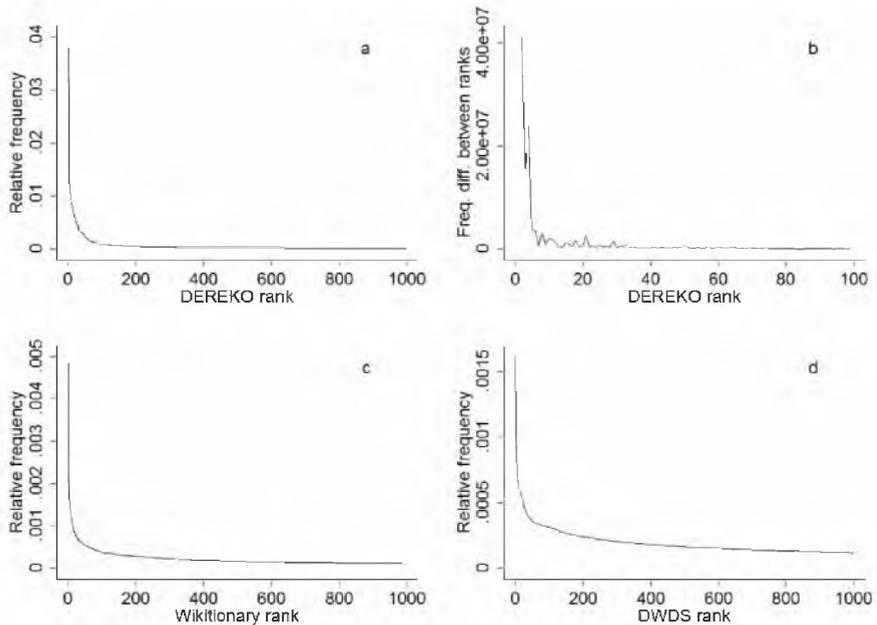


Fig. 1: Distributions of the corpus and the log file data. 1a: Relative frequency as a function of the DEREKO rank. 1b: Frequency difference between each successive rank as a function of the DEREKO rank. 1c/d: Relative frequency as a function of the Wiktionary/DWDS rank.

Category (<i>poms</i>)	Wiktionary log files (%)	DWDS log files (%)
1	57.94	57.30
2 - 10	33.71	31.15
11 - 49	6.69	9.09
50 - 500	1.63	2.44
500 +	0.03	0.02
Total	100.00 (abs. 185,071)	100.00 (abs. 156,478)

Tab. 1: Categorized relative frequency of the log file data.

The table shows two things: firstly, the Wiktionary and the DWDS log files are quite comparable on the *poms*-scale; secondly, just like the corpus data, the log files are heavily right skewed (cf. Figure 1c & Figure 1d). More than half of all query types

consist of phrases only searched for once *poms*. If we cumulate the first two categories, then we can state for both the Wiktionary and the DWDS data that 90% of the queries are requested 1 up to 10 times *poms*. So there is only a small fraction of all phrases in the log files that are searched for more frequently.

5 Analyzing the data

The problem

In the last section, we described the data and presented a new unit of measurement called *poms*. If we think about our research question again – whether dictionary users look up frequent words (frequently) – it is necessary to find an appropriate method for analyzing the data using this unit. For example, we could regress the log file frequency (in *poms*) on the corpus frequency, but an ordinary least squares (OLS) regression implies a linear relationship between the explanatory and the response variable, which is clearly not given. (Log-)Transforming both variables does not solve our problem, either, and this is in any case seldom a good strategy (O’Hara & Kotze, 2010). We could use the appropriate models for count data such as Poisson regression or negative binomial regression, but, as Baayen (2001, 2008, pp. 222–236) demonstrates at length, we still have to face the problem of a very large number of rare events (LNRE), which is typical for word frequency distributions. And even if we could fit such a model, it would remain far from clear what this would imply for our initial lexicographical question. Using the standard Pearson formula to correlate the corpus and the log file data suffers from the same nonlinearity problem as the OLS approach. Therefore de Schryver et al. (2006) implicitly used the nonparametric Spearman rank correlation coefficient which is essentially just the Pearson correlation between ranked variables. As mentioned above, we believe that this is still not the best solution, mainly because, on a conceptual level, ranking the corpus and log file data implies that subsequent ranks are equidistant in frequency, which is clearly not the case. Figure 1b plots the differences in frequency against the first 100 ranks for the DEREKO corpus data.

Again, the inherent Zipfian character of the distribution explains why the ranks are far from equidistant. For example, the difference in frequency between the first and the second rank is 251,480, whereas the difference between the 3000th and 3001th is only 5. Nevertheless the Spearman rank correlation coefficient treats the differences as equal.²³ The problem for data analysis becomes even more obvious

²³ In principle, we could use another similarity metric, for example the cosine measure (i.e. the normalized dot product, cf. Jurafsky & Martin, 2009, p. 699), but as in the case of using a count

when we tabulate categorized versions (described in the last section) of the data against each other (cf. Table 2, Table 3).

		DEREKO corpus rank		
		Top 200	rest	Total
Wiktionary logs	more than 10	86.50%	11.00%	11.17%
poms	rest	13.50%	89.00%	88.83%
	Total	100.00% (abs. 200)	100.00% (abs. 92,306)	100.00% (abs. 92,506)

Tab. 2: Crosstab of the DEREKO and the Wiktionary data ($X^2 = 1100.00$).

The tables reveal that of the top 200 DEREKO most frequent words, almost 90% are searched for more than 10 *poms* in Wiktionary or in DWDS. Because those 200 DEREKO word form types make up half of all tokens and because only about 10% of all phrases are searched for more than 10 *poms*, it seems that there is a relationship between corpus frequency and log file frequency. However, this relationship is far from linear.

		DEREKO corpus rank		
		Top 200	rest	Total
DWDS	more than 10	87.50%	15.77%	15.93%
logs poms	rest	12.50%	84.23%	84.07%
	Total	100.00% (abs. 200)	100.00% (abs. 92,306)	100.00% (abs. 92,506)

Tab. 3: Crosstab of the DEREKO and the DWDS data ($X^2 = 766.76$).

A possible solution

In the last section, we grouped the log files (cf. Table 1) into *poms* categories. We use this grouping again and stipulate the following categories: if a word form is searched for at least once *poms*, it is searched for *regularly*, if it is searched for at least twice, we call it *frequent*, and if it is searched for more than 10 times, it is *very frequent*. Table 4 sums up the resulting values. Please keep in mind that according to this definition, a *very frequent* search term also belongs to the *regular* and the *frequent* categories.

regression model, we are not sure what the value of the coefficient would actually imply both theoretically and practically.

Category	X searches <i>poms</i>	Wiktionary log files (%)	DWDS log files (%)
regular	at least 1	100.00	100.00
frequent	at least 2	42.06	42.70
very frequent	at least 11	8.35	11.55

Tab. 4: Definition of the categories used in the subsequent analysis and relative log file distribution.

Our definition is, of course, rather arbitrary, but due to the Zipf distribution of the data, only a minority of the searches (roughly 4 out of 10) occur more than once *poms* and even fewer words (roughly 1 out of 10) are searched for more than ten times *poms* (cf. Table 1). Therefore, this definition at least approximates the distribution of the log file data. Nevertheless, instead of using the categories presented in the first column in Table 4, we could also use the second column to label the categories, so it must be borne in mind that the labels merely have an illustrative function.

To solve the problem discussed above, we wrote a Stata program²⁴ that starts with the first ten DEREKO ranks and then increases the included ranks one rank at a time. At every step, the program calculates how many of the included word forms appear in the DWDS and Wiktionary log files *regularly*, *frequently*, and *very frequently* (scaled to percentage). Table 5 summarizes the results for 6 data points.

Included DEREKO ranks	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.0	87.5	99.5	99.5	86.5
2,000	96.9	91.0	67.6	98.4	96.0	64.9
10,000	85.5	72.9	47.5	86.3	75.3	40.2
15,000	80.3	66.5	41.8	77.4	66.1	33.7
30,000	69.4	54.6	31.3	62.7	50.9	23.4

Tab. 5: Relationship between corpus rank and log file data.

In this table, the relationship between the corpus rank and the log file data becomes obvious: the more DEREKO ranks we include, the smaller the percentage of those word forms appearing regularly/frequently/very frequently in both the DWDS and the Wiktionary log files. Let us assume for example that we prepare a dictionary of

²⁴ All Stata do files can be obtained upon request from AK (kopenig@ids-mannheim), who would also be happy to discuss any further technical or methodological details regarding this approach.

the 2,000 most frequent DEREKO word forms; our analysis of the DWDS and the Wiktionary data tells us that 96.9 % of those word forms are searched for *regularly* in DWDS, 91.0 % are searched for *frequently* and 66.6 % are searched for *very frequently*. For Wiktionary, these figures are a bit smaller (cf. Section 6.3 for a possible explanation).

Figure 2 plots this result for the DWDS and the Wiktionary log files separately. It comes as no surprise that the curve is different for the three categories, being steepest for the *very frequent* category, since this type of log file data only makes up a small fraction of the data (cf. Table 1 & Table 4).

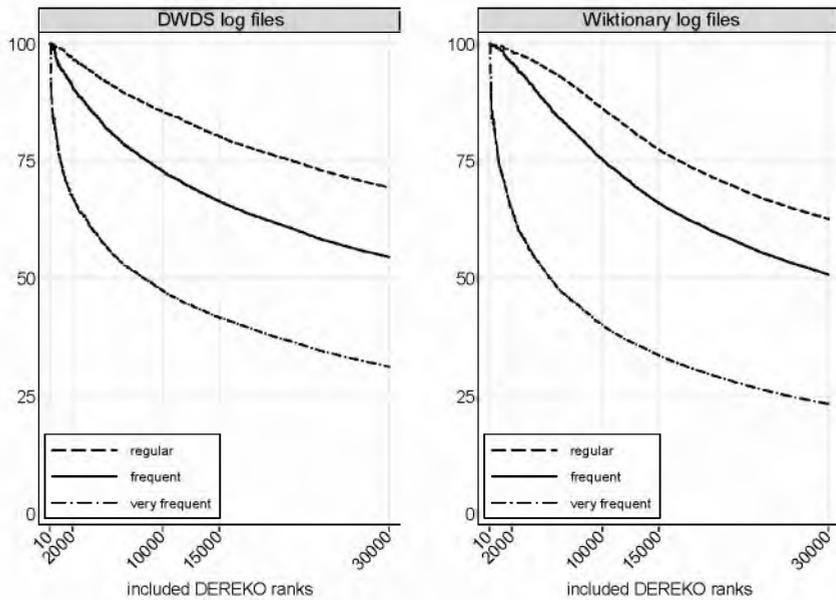


Fig. 2: Percentage of search requests which appear in the DWDS/Wiktionary log files as a function of the DEREKO rank.

Improving the solution

To further improve our analysis approach, we looked at the word forms that are absent in both the DWDS and the Wiktionary log files but that are present in the *unlemmatised* DEREKO corpus data. There is a roughly 60% overlap, which means that 6 out of ten word forms missing in the DWDS data are also missing in the

Wiktionary data. To understand this remarkable figure, we tried to find out more about the words that are missing in the log files but are present in the corpus data. Therefore, we used the Wiktionary word class information described in Section 2. Table 6 shows the information we gathered. For roughly 60% of the DEREKO word forms (that were absent in both the Wiktionary and the DWDS log files), no information was available in Wiktionary regarding word class. Table 6 also reveals that 15.52 % (last column) of the missing word forms belong to word classes that would not typically be found in a general (non-specialized) dictionary, i.e. declined and conjugated forms, toponyms and proper nouns.²⁵

Word class	Frequency	Relative frequency	Cumulative frequency
Declined form	10,168	10.99	10.99
Conjugated form	2,414	2.61	13.60
Toponym	977	1.06	14.66
Proper noun	793	0.86	15.52
Noun	14,366	15.53	31.05
Verb	2,442	2.64	33.69
Adjective	2,309	2.50	36.19
Partizip II (past participle)	785	0.85	37.04
Abbreviation	548	0.59	37.63
Adverb	463	0.50	38.13
Partizip I (participle)	91	0.10	38.23
Preposition	45	0.05	38.28
Other word classes/mixed cases	1,317	1.42	39.70
No information	55,788	60.31	100.00

Tab. 6: Wiktionary word form information about word forms that are present in the DEREKO corpus data but are absent in both the Wiktionary and the DWDS log files.

We then decided to rerun our analysis without these four word classes (printed in boldface in Table 6) and compare the initial results with the updated ones. Table 7 again summarizes the results for 6 data points, while Figure 3 superimposes the updated results of Figure 3 with the initial results coloured in light-grey. For example, our results show that if we prepared a dictionary with the 15,000 most frequent

²⁵ There are of course mixed cases in the Wiktionary word class information data because a word can have multiple meanings. For example, “Hirsch” (stag) can either be a common noun or a family name. In all those cases, we did not exclude those words from the subsequent analysis.

DEREKO word forms, all of those word forms are looked up in the DWDS and the Wiktionary on a regular basis, 83.3%/90.4% are looked up frequently in the DWDS/Wiktionary and roughly half of those words forms are looked up very frequently in both the DWDS and Wiktionary.

Included DEREKO ranks	DWDS (%)			Wiktionary (%)		
	regular	frequent	very frequent	regular	frequent	very frequent
10	100.0	100.0	100.0	100.0	100.0	100.0
200	100.0	99.5	95.5	100.0	100.0	98.0
2,000	100.0	96.7	84.8	100.0	98.9	80.1
10,000	100.0	86.8	62.3	100.0	92.8	54.6
15,000	100.0	83.3	54.7	100.0	90.4	47.0
30,000	100.0	77.4	40.6	86.2	75.1	32.1

Tab. 7: Relationship between corpus rank and log file data (updated data).

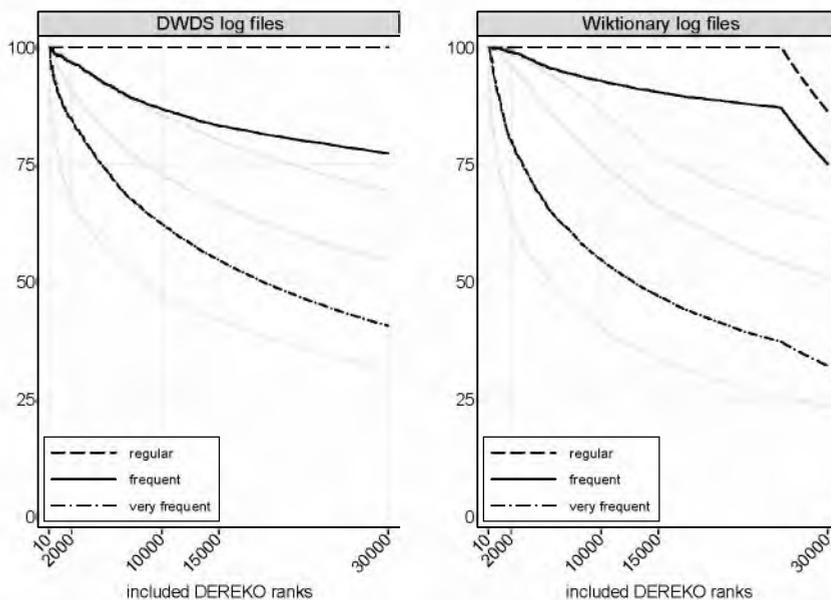


Fig. 3: Percentage of search requests appearing in the DWDS/Wiktionary log files as a function of the DEREKO rank (updated data in black, original data in grey).

It is rather unsurprising that this step considerably improves our initial results because – like de Schryver et al. (2006) – we used an unlemmatized word list. So in general, our results seem to suggest that it makes more sense to use a lemmatized version of the corpus word list. To check this, we used a lemmatized DEREKO word list.²⁶ Figure 4 shows that our assumption seems to be correct as the results are better for the lemmatized list compared to the unlemmatized list, especially for the DWDS data.

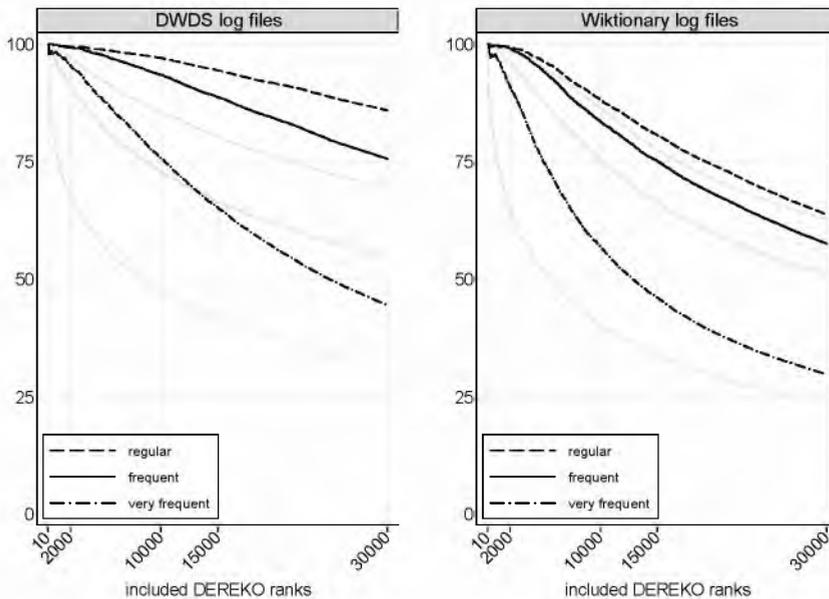


Fig. 4: Percentage of search requests appearing in the DWDS/Wiktionary log files as a function of the DEREKO rank (lemmatized data in black, unlemmatized data in grey).

Evaluating the results

Before we discuss our results further in the conclusion, we would like to provide an additional impression of our results by asking what proportion of all search requests (tokens) could be covered with such a corpus-based strategy. Table 8 shows the

²⁶ Again, we used the most recent version of this list published in December 2012 available here <http://www1.ids-mannheim.de/kl/projekte/methoden/derewo.html> (last accessed 20 June 2013), which we slightly modified in a rough-and-ready manner.

percentage of all logged search request tokens that would be successful if the first X DEREKO ranks (first column; again with the unlemmatized but updated DEREKO list, cf. Section 6.3) were entered into the relevant dictionary for the DWDS and the Wiktionary data separately. If we again use the example of the first 15,000 DEREKO most frequent word forms, then around half of all DWDS search requests that occur regularly or frequently (*poms*) are covered, while around two-thirds of all very frequent requests are successful. If we included the 30,000 most frequent DEREKO words, roughly two-thirds of the *regular* and *frequent* and 80.0% of the *very frequent* DWDS search requests would be covered in the dictionary. In other words, this means if we included the 30,000 most frequent DEREKO word forms, the vast majority of requests would be successful. In general, these figures are smaller for the Wiktionary data. Why is that the case? If we look at the data, we see that in Wiktionary, many users search for abbreviations. For example, of the 50 most frequent queries, six are word forms abbreviating typical internet slang phrases (“www”, “wtf”, “imho”, “lmao”, “afk”, “lol”, “aka”), and these make up 12.6 % of all the first 50 query tokens. If we use Google to find out what those abbreviations mean, in 4 out of those 6 cases, the first result presented is a link to Wiktionary; in one case (“lol”), a Wiktionary link is listed under the top 5 hits.

Included DEREKO ranks	Percentage of all DWDS log tokens			Percentage of all Wiktionary log tokens		
	regular	frequent	very frequent	regular	frequent	very frequent
10	0.2	0.3	0.3	0.1	0.1	0.1
200	3.8	4.1	5.2	1.8	2.0	2.7
2,000	19.3	21.2	26.5	9.8	11.0	14.7
10,000	42.4	46.4	56.7	25.9	29.0	36.4
15,000	49.8	54.4	65.7	34.1	38.1	46.7
30,000	63.7	69.2	80.0	49.3	54.9	64.8

Tab. 8: Percentage of log file data covered as a function of the DEREKO rank.

Conclusion

In general, the use of a corpus for linguistic purposes is based on one assumption:

“It is common practice of corpus linguistics to assume that the frequency distributions of tokens and types of linguistic phenomena in corpora have - to put it as generally as possible - some kind of significance. Essentially more frequently occurring structures are believed to hold a more prominent place, not only in actual discourse but also in the linguistic system, than those occurring less often.” (Schmid, 2010, p. 101)

We hope that we have provided evidence in this chapter which shows that, based on this assumption, corpus information can also be used fruitfully when it comes to deciding which words to include in a dictionary.²⁷

If we think about our fictional word “bfk”, which we used as an example in the introduction, most probably everyone will agree that the corpus indeed tells us that it is better to exclude this word from any dictionary. Nevertheless, de Schryver et al. (2006, pp. 78–79) conclude their study by saying that:

“[T]he corpus does not provide the ‘magic answer’ every dictionary maker was hoping for [...] There is thus no such thing as words a lexicographer better not treat.”

While we agree that a corpus-based strategy is not the “magic answer”, we simply think it is the best one there is, if the aim of a lexicographical project is either to provide a general description of the vocabulary, or to compile a specialized dictionary for a particular user group. In both cases, a balanced or a special corpus can help to select entries in an economical and intersubjectively traceable manner. Are there any other systematic alternatives? If we again consult the OED frequently asked questions, we find how it used to be before large collections of texts illustrating actual language use were available:

“In previous centuries dictionaries tended to contain lists of words that their writers thought might be useful, even if there was no evidence that anyone had ever actually used these words.”²⁸

Exactly this evidence can be found in a corpus and our analysis shows that the frequency information can serve as a proxy for the lookup probability in a dictionary. Maybe one last analysis will drive home our point: if it really does not make any difference which words are included in a dictionary “beyond the top few thousand words” as de Schryver et al. put it (2006, p. 79), then we can drop the 10,000 most frequent DEREKO word forms and then just randomly sample 10,000 of the remaining word forms for our dictionary. If we calculate how many of those word forms are actually being looked up, we find that for the Wiktionary data 34 % and for the DWDS data 45 % of the described word forms are actually being looked up at least once per one million search requests. What happens if we instead base our dictionary on the corpus frequency and describe rank 10,001 up to rank 20,000 in our hypothetical dictionary? In that case, for the Wiktionary data 56% (instead of 34%)

²⁷ It is interesting to note that although the DWDS log files are actual search requests, while the Wiktionary data consist of page views (as mentioned in Section 3), the results for both dictionaries point in the same direction.

²⁸ <http://oxforddictionaries.com/words/how-do-you-decide-whether-a-new-word-should-be-included-in-an-oxford-dictionary> (last accessed 20 June 2013).

and for the DWDS data 67% (instead of 45%)²⁹ are actually being looked up at least once per one million search requests. In a nutshell: our results imply that dictionary users do look up frequent words.

Bibliography

- Baayen, R. H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.
- Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge, UK: Cambridge University Press.
- De Schryver, G.-M., Joffe, D., Joffe, P., & Hillewaert, S. (2006). Do dictionary users really look up frequent words?—on the overestimation of the value of corpus-based lexicography. *Lexikos*, 16, 67–83.
- Hanks, P. (2012). Corpus evidence and electronic lexicography. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 57–82). Oxford: Oxford University Press.
- Jurafsky, D., & Martin, J. H. (2009). *Speech and Language processing: an introduction to natural language processing, computational Linguistics, and speech recognition*. Upper Saddle River: Pearson Education (US).
- Kupietz, M., Belica, C., Keibel, H., & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, D. Tapias, M. Rosner, S. Piperidis, J. Odjik, J. Mariani, ... K. Choukri (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-10)* (pp. 1848–1854). Valetta, Malta: European Language Resources Association (ELRA).
- Ludwig-Mayerhofer, W. (2011). Ilmes – Internet Lexikon der Methoden der empirischen Sozialforschung. *ILMES – Internet-Lexikon der Methoden der empirischen Sozialforschung*. Retrieved September 14, 2013, from <http://www.lrz.de/~wlm/ilmes.htm>
- Meyer, C. M., & Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 259–291). Oxford: Oxford University Press.
- Nesi, H. (2012). Alternative e-dictionaries: Uncovering dark practices. In S. Granger & M. Paquot (Eds.), *Electronic lexicography* (pp. 363–378). Oxford: Oxford University Press.
- O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2), 118–112.
- Rundell, M. (2012). It works in practice but will it work in theory? The uneasy relationship between lexicography and matters theoretical. In J. M. Torjusen & R. V. Fjeld (Eds.), *Proceedings of the 15th EURALEX International Congress 2012, Oslo, Norway, 7 – 11 August 2012*. Oslo. Retrieved September 14, 2013, from http://www.euralex.org/elx_proceedings/Euralex2012/pp47-92%20Rundell.pdf
- Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101–133). Berlin, New York: de Gruyter.

²⁹ In other words, the corpus based-strategy improves the rate of success by roughly 22 percentage points for both the DWDS and the Wiktionary data.

Verlinde, S., & Binon, J. (2010). Monitoring Dictionary Use in the Electronic Age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Ljouwert: Afûk.

